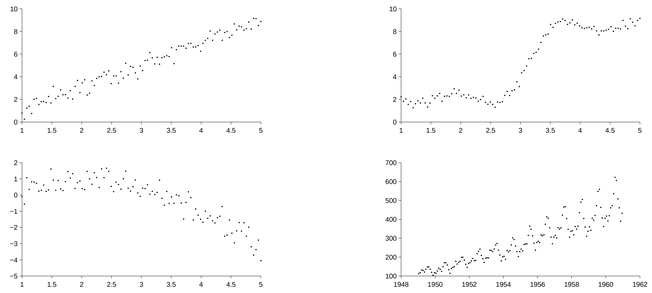
Automating pattern discovery and the statistical process for regression

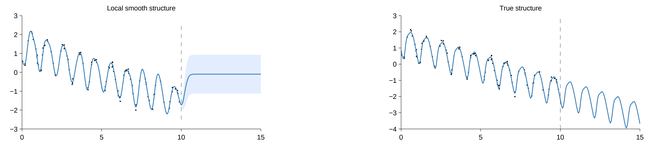
David Duvenaud[1], **James Robert Lloyd**[1], Roger Grosse[2],
Joshua B. Tenenbaum[2], Zoubin Ghahramani[1]

1: Department of Engineering, University of Cambridge, UK 2: Massachusetts Institute of Technology, USA

Data often exhibits high level structure or patterns



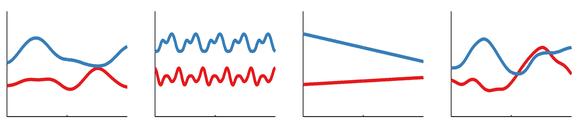
Identifying this structure is crucial for extrapolation



- Traditionally, a researcher / scientist / statistician would select an appropriate model for the type of structures present
- Automatic model selection techniques already exist, typically choosing between a finite or restricted set of models
- Instead, we automate statistical model *construction*

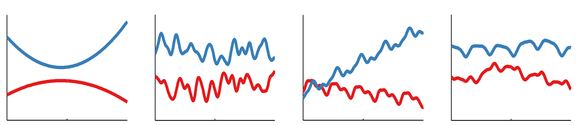
Gaussian process regression can model many structures with an appropriately chosen kernel

- The kernel encodes the inductive bias of the model i.e. the types of functions the model 'believes in'
- Below we list standard base kernels, and examples of functions the model believes in (samples from the prior)



SE	Per	Lin	RQ
local variation	repeating structure	linear functions	multi-scale variation

- Base kernels can be combined to create more complicated structural assumptions



Lin × Lin	SE × Per	Lin + Per	SE + Per
quadratic functions	locally periodic	periodic with trend	periodic with noise

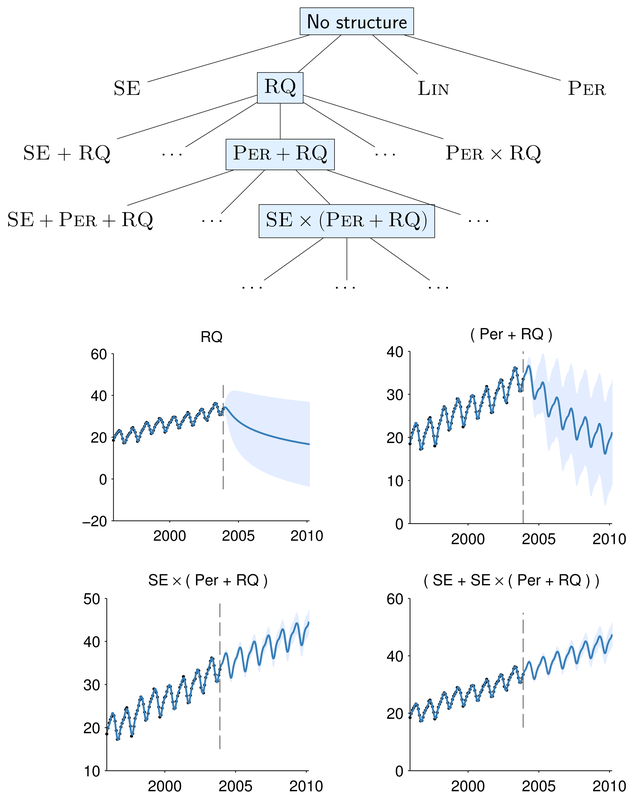
We consider all kernel expressions derived from a generative grammar. . .

- Constructing appropriate composite kernels has previously been the domain of Gaussian process experts
- We consider all algebraic expressions involving a small number of base kernels and the operations '+' and '×', including e.g.

Bayesian linear regression	$\textsc{Lin}$
Bayesian polynomial regression	$\textsc{Lin} \times \textsc{Lin} \times \ldots$
Generalized Fourier decomposition	$\textsc{Per} + \textsc{Per} + \ldots$
Generalized additive models	$\sum_{d=1}^{D} \mathrm{SE}_d$
Automatic relevance determination	$\prod_{d=1}^{D} \mathrm{SE}_d$
Linear trend with deviations	$\textsc{Lin} + \mathrm{SE}$
Linearly growing amplitude	$\textsc{Lin} \times \mathrm{SE}$

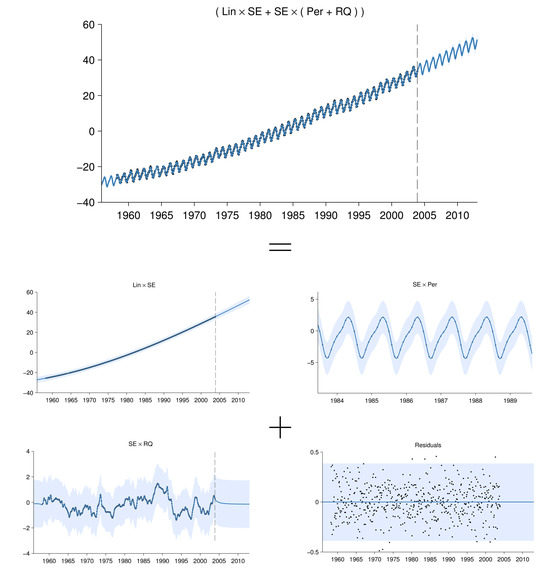
. . . which we search greedily, producing progressively better statistical models

- We try all base kernels, selecting the one with the highest (approximate) marginal likelihood which balances data fit and model complexity
- The search continues by adding an extra term to the current best kernel, stopping when marginal likelihood no longer improves



Example: Mauna Loa CO_2 concentration

- By automatically inferring an appropriate kernel, we can also automatically decompose functions into additive components (additive components of the kernel correspond to independent additive functions)



Example: International airline passengers